PERMUTATION VERSUS BOOTSTRAP SIGNIFICANCE TESTS IN MULTIPLE REGRESSION AND ANOVA

Cajo J. F. ter Braak
Agricultural Mathematics Group
Box 100, 6700 AC Wageningen, the Netherlands

Kempthorne's (1952) formulation of the randomization test is extended to yield a permutational analog of the bootstrap significance test. In the new test, residuals of a multiple regression are permuted instead of being bootstrapped. The test is an attractive alternative for Oja's test that permutes predictors (Austr. J. Statist. 29, 91-100, 1987).

1. Introduction

Permutation tests have a long history dating back at least to Neyman (1923). As there were no computers at that time, permutation tests were not much used, but served as an additional rationale for the validity of the ANOVA F-test. Pitman (1937) and Kempthorne (1952) showed that, in simple cases, the F-test yielded an approximation to the full permutation test. Permutation tests have a simple basis for their validity: randomization in experimental design (Kempthorne 1952, Scheffé 1959) and the exchangeability assumption in observational studies. Permutation tests and related nonparametric tests based on ranks (Cox & Hinkley 1974, Lehmann 1975) exist only for simple cases and for these cases they give exact significance levels. Their application is, however, problematic both in regression (Brown & Maritz 1982, Oja 1987) and in more complex ANOVA situations, such as testing for interaction in factorial experiments. Despite the existence of Puri & Sen's (1971: section 7.7) rank test for interaction, Edgington (1987: 144) claimed that there is no valid permutation test for interaction, because "This H_0 does not refer to measurements for individual subjects, and so there is no basis for generating data permutations for alternative subject assignments, a necessity for determining significance by data permutation".

K.-H. Jöckel et al. (eds.), *Bootstrapping and Related Techniques*

The bootstrap is relatively recent (Efron, 1979). Its main strength lies in estimating standard errors and in constructing confidence intervals (DiCiccio & Romano 1988). Bootstrap significance tests can in principle be derived from confidence intervals, but a direct approach also exists (Beran 1986, 1988, Jöckel 1990). Compared to permutation tests, the bootstrap significance test can be applied to much more complex problems, but despite the nice asymptotic properties proved by Hall & Titterington (1989) its validity in small samples is more speculative (Romano 1988, 1989).

In this paper, permutation tests are compared with bootstrap tests for use in (multivariate) regression and ANOVA. In both approaches, there is the choice what to permute or to bootstrap: the data values themselves or the residuals? In Kempthorne's (1952) formulation, the permutation test uses randomized "fixed plot errors". With errors replaced by residuals, Kempthorne's permutation test is shown to be very similar to the bootstrap test that uses residuals, the only difference being whether residuals are sampled with or without replacement. The new permutation test is briefly compared with Oja's (1987) permutation test that permutes predictors.

2. Bootstrap significance test

We consider the regression model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon \tag{2.1}$$

where $\mathbf{y}$ is a random n-vector of responses on n units, $\mathbf{X}$ and $\mathbf{Z}$ are fixed known $n\times p$ and $n\times q$ matrices, the columns of which contain explanatory variables, β and γ are p- and q-vectors of unknown, fixed regression coefficients and ε contains random errors with zero mean and unknown constant variance σ^2. Our interest focuses on the effect of the variables in $\mathbf{X}$ on $\mathbf{y}$ in the presence of the covariables in $\mathbf{Z}$. The parameter of interest is thus β, whereas γ and σ are nuisance parameters. In particular, we wish to test the hypothesis $\beta=\beta_0$ using the F-ratio.

Efron (1982) proposes two bootstrap schemes for regression. The one-sample bootstrap that resamples the statistical units has no permutational analog and will therefore not be considered. The other scheme resamples the observed residuals. A Monte Carlo bootstrap significance test is obtained as follows (Hall & Titterington 1989):

1. Regress $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{Z}$, yielding estimates $\mathbf{b}$ and $\mathbf{c}$ for β and γ, and from these, fitted values $\hat{\mathbf{y}}=\mathbf{Xb}+\mathbf{Zc}$ and residuals $\mathbf{e}=\mathbf{y}-\hat{\mathbf{y}}$.
2. Calculate the F-ratio, F_{obs} say, based on $\mathbf{y}$ for testing the hypothesis $\beta=\beta_0$.
3. Draw a bootstrap sample $\mathbf{e}^*$ from $\mathbf{e}$ and calculate

$$\mathbf{y}^* = \mathbf{Xb} + \mathbf{Zc} + \mathbf{e}^* \qquad (2.2)$$

4. Regress $\mathbf{y}^*$ on $\mathbf{X}$ and $\mathbf{Z}$, yielding estimates $\mathbf{b}^*$ and $\mathbf{c}^*$.
5. Calculate the F-ratio, F^* say, based on $\mathbf{y}^*$ for testing the hypothesis $\beta=\mathbf{b}$.
6. Repeat steps 3-5 until m bootstrap values of F^* are calculated.
7. The bootstrap estimate of the significance level is $k/(m+1)$, where k is the rank of F_{obs} among the bootstrap values of F^* when ranked from high to low.

The method works, at least in large samples, because the variability in $\mathbf{b}$ around the true value β is mimicked by the variability of $\mathbf{b}^*$ around the true value $\mathbf{b}$ in the bootstrap samples. Similarly, the variability in F_{obs} for testing $\beta=\beta_0$ is mimicked by the variability in F^* for testing $\beta=\mathbf{b}$. The mimicking is good because the test statistic is asymptotically pivotal (Hall & Titterington 1989): for normal errors, the distribution of the F-ratio does not depend on β, γ and σ and for nonnormal errors this will be asymptotically true. The procedure uses residuals under the alternative hypothesis (value of β unconstrained) to increase the power compared to a procedure using residuals from the null model (Hall & Titterington 1989).

The calculation of F^* can be simplified. Step 4 uses model (2.1) with $\mathbf{y}^*$ replacing $\mathbf{y}$. Therefore,

$$\mathbf{e}^* = \mathbf{y}^* - \hat{\mathbf{y}} = \mathbf{X}(\beta-\mathbf{b}) + \mathbf{Z}(\gamma-\mathbf{c}) + \varepsilon \qquad (2.3)$$

F^* in step 5 can thus be obtained from testing the hypothesis $\beta-\mathbf{b}=\mathbf{0}$ in (2.3), i.e. by regressing $\mathbf{e}^*$ on $\mathbf{Z}$ and then adding $\mathbf{X}$ to the regression, giving residual sums of squares RSS_Z and RSS_{X+Z}, respectively. Then,

$$F^* = \{ (RSS_Z - RSS_{X+Z}) / p \} / \{ RSS_{X+Z} / (n-p-q) \} \qquad (2.4)$$

3. Permutation tests

For investigating the validity of the F-test in the analysis of randomized block experiments, Kempthorne (1952: section 8.2) proposed a randomization model in which a fixed plot error is randomly assigned to treatments. He writes

"If we denote the observed yield of treatment k in block i by y_{ik}, we may write

$$y_{ik} = \mu + b_i + t_k + \sum_j \delta_{ij}^k e_{ij} \qquad (3.1)$$

where δ_{ij}^k is equal to unity if treatment k occurs on plot j in the ith block and is zero otherwise. The random error attached to any observed yield is the whole expression $\sum_j \delta_{ij}^k e_{ij}$. Any particular e_{ij} is a fixed variable which we do not know. The random variable in the expression (3.1) is the term δ_{ij}^k, and its distribution is determined by the randomization procedure which is used in obtaining the particular experimental plan."

Notice that the random error in (3.1) specifies a permutation of the fixed plot errors within each block. With complete randomization, the model entails that we could equally well have observed, using our regression notation,

$$\mathbf{y}^+ = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon^+ \qquad (3.2)$$

where ε^+ denotes a permutation of the error vector ε. By replacing the unknown β,γ and ε by their estimates, we obtain the permutational analog of (2.2)

$$\mathbf{y}^+ = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{c} + \mathbf{e}^+ \qquad (3.3)$$

yielding estimates $\mathbf{b}^+$ and $\mathbf{c}^+$. We consider this analog further in section 4.

In developing a permutation test for use in multiple regression and covariance analysis, Oja (1987) rejected (3.2) on the basis that the errors in (3.2) are unknown, but required for his test statistic. Instead, he stressed that randomization randomly assigns treatment values to units, so that the model becomes

$$\mathbf{y} = \mathbf{X}^+\beta + \mathbf{Z}\gamma + \varepsilon \qquad (3.4)$$

where $\mathbf{X}^+$ is obtained by permuting the rows of the matrix $\mathbf{X}$ (see also Collins 1987). This model underlays the permutation test in the computer program CANOCO version 2.1 (ter Braak 1988a,c). For use in observational studies, this model can be motivated by assuming exchangeability of the rows of $\mathbf{X}$.

If permutations are carried out under the null hypothesis $\beta=\beta_0=\mathbf{0}$ (as is usual), then the distinction between (3.2) and (3.4) vanishes if $\gamma=\mathbf{0}$ and also if permutations are restricted within blocks, as in Kempthorne (1952). Then, it is immaterial whether $\mathbf{y}$ or the rows of $\mathbf{X}$ or the residuals $\mathbf{y}$-$\hat{\mathbf{y}}$ are permuted.

Model (3.4) is perfect for covariance analysis applied to randomized experiments where the covariates $\mathbf{Z}$ are measured before the random assignment of treatments to experimental units: the covariates are fixed and permutation of the rows of $\mathbf{X}$ mimics the randomization at the design stage. But, model (3.4) has a disadvantage with multiple regression: each permutation modifies the design matrix. Therefore, with each permutation the F-ratio measures the effect of another contrast (specified by the

projection of $\mathbf{X}^+$ on the orthocomplement of $\mathbf{Z}$). To say the same more theoretically, model (3.4) violates the principle that the design is ancillary (Welch 1990), whereas model (3.3) obeys the ancillarity principle. Moreover, (3.4) does not allow for a permutation test of interaction[1], whereas (3.3) does.

4. Bootstrap versus permutation

In this section, the bootstrap (2.2) is compared with its permutational analog (3.3). In the bootstrap (2.2), residuals are resampled with replacement, whereas in permutation (3.3) the same residuals are resampled without replacement, resulting in correlation $-(n\text{-}1)^{-1}$ between the errors (Lehmann 1975: (A.41)). Because the error covariance matrices of $\mathbf{e}^*$ and $\mathbf{e}^+$ are known, standard linear model theory (Seber 1977: section 3.7.1) can be used to obtain the expected value and the variance of $\mathbf{b}^*$ and $\mathbf{b}^+$. The essential formulae were given by Efron (1982: example 5.6) for the bootstrap and by Cox & Hinkley (1974: example 6.2) and Lehmann (1975: (A48-49)) for permutation. The comparison is simplified if we assume that the constant vector is one of the columns in $\mathbf{Z}$. Excluding regression through the origin, we obtain

$$E^*(\mathbf{b}^*) = E^+(\mathbf{b}^+) = \mathbf{b} \tag{4.1}$$

$$\text{var}^*(\mathbf{b}^*) = (1 - 1/n)\ \text{var}^+(\mathbf{b}^+) = s^2\ (\mathring{\mathbf{X}}'\mathring{\mathbf{X}})^{-1} \tag{4.2}$$

where s^2 is the mean square of the residuals $\mathbf{e}$ and $\mathring{\mathbf{X}}$ is obtained by projection of $\mathbf{X}$ on the orthocomplement of $\mathbf{Z}$. Second and higher order moments of $\mathbf{b}^*$ and $\mathbf{b}^+$ differ by $O(1/n)$. Equations (4.1-2) plus this order property form the justification for proposing the permutation model as alternative for the bootstrap model. It remains to be studied how the distributions of F^* and F^+ differ. Using permutations and bootstraps under the null hypothesis, Romano (1989) proved that the corresponding tests yield asymptotically the same critical values under quite general conditions. His results can probably be extended to our tests.

---

[1] For example, when the columns $\mathbf{z}_1$ and $\mathbf{z}_2$ are kept fixed, permutation of the column $\mathbf{x}_{12}$ which contains their elementwise product, yields an unfeasible design: $\mathbf{x}^+_{12}$ is not an interaction term.

5. Discussion

Traditionally, permutation is always performed under the null hypothesis and then only if the residuals are exchangeable (e.g. Puri & Sen 1971, Romano 1989, Welch 1990). Our results (4.1-2) show the validity of permutation under the alternative hypothesis, even if residuals are not exchangeable because of differences in their variances. It is thus not needed to studentize residuals. If, however, the variance of ε is heterogeneous, this knowledge can be used to standardize the residuals (Hinkley 1988: 331). In the program CANOCO version 3.1 (ter Braak 1988a,c) this device is used to obtain permutation tests for use in partial canonical correspondence analysis (ter Braak 1988b).

Freedman & Lane (1983) used a model similar to (3.3). Their proposal differs from (3.3) in that $\mathbf{e}$ is estimated under the null hypothesis $\beta=\beta_0$. But, the expectation and variance of $\mathbf{b}^+$ have the classical form (4.1-2) only when using $\mathbf{e}$ from the alternative hypothesis. By contrast, in developing a permutation test for generalized linear models, Gail et al (1988) used Oja's (1987) approach (3.4) of permuting the treatment variables.

In the traditional permutation tests, the test statistic can often be simplified, e.g. from an F-ratio to a sum of squares (Edgington 1987). Using (3.3), there is no such simplication. Hall & Titterington (1989) show the gain in level accuracy by using an asymptotically pivotal statistic such as the F-ratio.

If randomization is used to obtain the data, reference to an underlying population model can be avoided by using the permutation approach (cf Pitman 1937, Cox & Hinkley 1974) instead of the bootstrap. In principle, the randomization model, whence permutation, also forms a basis for constructing confidence regions.

In practice, bootstrap and permutation are often carried out by Monte Carlo methods. Balanced bootstraps and other variance reduction techniques may then help to increase the power when only a limited number of bootstrap samples can be made (Hinkley 1988). Permutation may have some advantage here because there is maximum balance in each random permutation.

Notice that the set of all permutations is a proper subset of the set of all bootstrap samples. As a result, permutation is sometimes not applicable, e.g. in one-sample problems where all permutations yield the same mean value. For the rest, bootstrap and permutation tests are just different but related tests.

6. References

Beran, R. (1986). Simulated power functions. *Ann. Statist.* **14,** 151-173.

Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *J. Amer. Statist.* **83,** 687-697.

Brown, B. M. and Maritz, J. S. (1982). Distribution-free methods in regression. *Austral. J. Statist.* **24,** 318-331.

Collins, M. F. (1987). A permutation test for planar regression. *Austral. J. Statist.* **29,** 303-308.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

DiCiccio, T. J. and Romano, J. P. (1989). A review of bootstrap confidence intervals. *J. R. Statist. Soc. B* **50,** 338-354.

Edgington, E. S. (1987). *Randomization Tests. 2nd Ed.* New York: Dekker.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7,** 1-26.

Efron, B. (1982). *The Jackknife, the Bootstrap and other Resampling Plans.* Philadelphia: SIAM.

Freedman, D. A. and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *J. Bus. Econ. Statist.* **1,** 292-298.

Gail, M. H., Tan, W. Y. and Piantadosi, S. (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika* **75,** 57-64.

Hall, P. and Titterington, D. M. (1989). The effect of simulation order on level accuracy and power of Monte Carlo tests. *J. R. Statist. Soc. B* **51,** 459-467.

Hinkley, D. V. (1988). Bootstap methods. *J. R. Statist. Soc. B* **50,** 321-337.

Jöckel, K.-H. (1990). Monte Carlo techniques and hypothesis testing. In *1rst Int. conf. Statistical computing, Cesme, Izmir 1987,* E. J. Dudewisz (ed.).

Kempthorne, O. (1952). *The Design and Analysis of Experiments.* New York: Wiley.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: principles (in Polish with German summary). *Roczniki Nauk Rolniczch* **10,** 21-51.

Oja, H. (1987). On permutation tests in multiple regression and analysis of covariance analysis problems. *Austral. J. Statist.* **29,** 91-100.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. III The analysis of variance test. *Biometrika* **29,** 322-335.

Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis.* New York: Wiley.

Romano, J. P. (1988). A bootstrap revival of some nonparametric distance tests. *J. Amer. Statist.* **83,** 698-708.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.* **17,** 141-159.

Scheffé, H. (1959). *The Analysis of Variance.* New York: Wiley.

Seber, G. A. F. (1977). *Linear Regression Analysis.* New York: Wiley.

ter Braak, C. J. F. (1988a). CANOCO - an extension of DECORANA to analyze species-environment relationships. *Vegetatio* **75,** 159-160.

ter Braak, C. J. F. (1988b). Partial canonical correspondence analysis. In *Classification and Related Methods of Data Analysis,* H. H. Bock (ed.), 551-558. Amsterdam: North-Holland.

ter Braak, C. J. F. (1988c). *CANOCO - a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1). Report LWA-88-02.* Wageningen: Agricultural Mathematics Group.

Welch, W. J. (1990). Construction of permutation tests. *J. Amer. Statist.* **85,** 693-698.